A test of independence based on a sign covariance related to Kendall's tau

Wicher Bergsma*and Angelos Dassios
London School of Economics and Political Science

January 25, 2011

Abstract

The standard two-variable chi-square test is typically consistent for all alternatives to independence, but effectively treats the data as nominal which may lead to loss of power for ordinal data. Alternatively, a test based on Kendall's tau does take ordinality into account, but only has power against a narrow set of alternatives. This paper introduces a new test aimed at filling this gap, i.e., it is designed for ordinal data and to have asymptotic power for all alternatives. Our test is a permutation test based on a modification of Kendall's tau, denoted τ^*, which is nonnegative and equal to zero if and only if and only if independence holds. An interpretation of τ^* in terms of concordance and discordance for sets of four observations is given. The new coefficient is a sign version of a covariance introduced by Bergsma (2006).

Keywords: measure of association, test of independence, concordance, discordance, sign test, ordinal data, permutation test, copula.

1 Introduction and overview of main results

Sign covariances such as Kendall's tau (τ) are especially useful for testing independence when (i) the data are ordinal (whether continuous or discontinuous) and the ordinary covariance is inappropriate, (ii) the data are heavy tailed and the ordinary covariance may not be defined, or (iii) the data are contaminated and robustness is needed. Kendall's tau (τ), however, has the possible drawback that it may be zero when there is association present. To achieve power against broader alternatives, the chi-square test can be used; it is directly applicable to categorical data and can be used for continuous data after a suitable categorization. However, the chi-square test for data with ordered outcomes does not take the ordinal nature of the data into account, leading to potential power loss for 'ordinal' alternatives; effectively the chi-square test treats the data as nominal rather than ordinal (see also Agresti, 2010). As a possible substitute for these two tests, we consider a modification of τ^2, which we call τ^*. Tests based on the sample value of τ^* yield power against a broad range of ordinal alternatives. Below, we first derive τ^*, then summarize its properties and provide a probabilistic interpretation in terms of concordance and discordance probabilities.

Regarding notation, we denote iid sample values by $(x_1, y_1), \ldots, (x_n, y_n)$, but will also use $\{(X_i, Y_i)\}$ to denote iid replications of (X, Y) in order to define population coefficients. For the

*Corresponding author

most part of this paper we assume all variables are real, but briefly touch upon more general metric sample spaces. The empirical value t of Kendall's tau is

$$t = \frac{1}{n^2} \sum_{i,j=1}^{n} \operatorname{sign}(x_i - x_j) \operatorname{sign}(y_i - y_j)$$

and its population version is

$$\tau = E \operatorname{sign}(X_1 - X_2) \operatorname{sign}(Y_1 - Y_2)$$

(Kruskal, 1958; Kendall & Gibbons, 1990). With

$$\begin{aligned} s(z_1, z_2, z_3, z_4) &= \operatorname{sign}(z_1 - z_3)(z_2 - z_4) \\ &= \operatorname{sign}(|z_1 - z_2|^2 + |z_3 - z_4|^2 - |z_1 - z_3|^2 - |z_2 - z_4|^2) \end{aligned}$$

we obtain

$$t^2 = \frac{1}{n^4} \sum_{i,j,k,l=1}^{n} s(x_i, x_j, x_k, x_l) s(y_i, y_j, y_k, y_l)$$

and

$$\tau^2 = E s(X_1, X_2, X_3, X_4) s(Y_1, Y_2, Y_3, Y_4)$$

Replacing squared differences in s by absolute values of differences, we define

$$a(z_1, z_2, z_3, z_4) = \operatorname{sign}\left(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|\right) \tag{1}$$

This leads to a modified version of t^2,

$$t^* = \sum_{i,j,k,l} a(x_i, x_j, x_k, x_l) a(y_i, y_j, y_k, y_l)$$

and the corresponding population coefficient

$$\tau^* = \tau^*(X, Y) = E a(X_1, X_2, X_3, X_4) a(Y_1, Y_2, Y_3, Y_4)$$

The quantities t^* and τ^* are new, and the main result of the paper is the following:

Theorem 1 *Let X and Y be real continuous random variables. Then $\tau^*(X, Y) \geq 0$ with equality if and only if X and Y are independent.*

The proof is given in Section 2. We conjecture that the continuity condition is not needed, and have the following partial proofs of this: (i) Lemma 1 in Section 3.1 shows that if X is binary and Y continuous or discrete, the theorem holds; (ii) the code for a computational 'proof' of nonnegativity of τ^* using *Mathematica* for 3×3 contingency tables with given marginals is given in Appendix B.

If the sign functions are omitted from τ^*, we obtain the covariance κ introduced by Bergsma (2006). He showed that for arbitrary real random variables X and Y, $\kappa(X, Y) \geq 0$ with equality if and only if X and Y are independent.

We now give a probabilistic interpretation of τ^*. A pair of points $\{(x_1, y_1), (x_2, y_2)\}$ is called concordant if $(x_1 - x_2)(y_1 - y_2) > 0$ and discordant if $(x_1 - x_2)(y_1 - y_2) < 0$, as illustrated in Figure 1. Denoting the probabilities that two randomly chosen points are concordant by Π_{C_2} and that they are discordant by Π_{D_2}, Kendall's tau has the well-known probabilistic interpretation

$$\tau = \Pi_{C_2} - \Pi_{D_2}$$

An analogous interpretation of τ^* can be given. A set of four points is concordant if there exist vertical and horizontal axes such that two opposing open quadrants contain two points each (Figure 2(a)). The set is discordant if there exist vertical and horizontal axes such that every open quadrant contains a single point (Figure 2(b)). Note that the axes must strictly separate the points, i.e., no points can fall on the axes. In mathematical notation, a set of four points $\{(x_1, y_1), \ldots, (x_4, y_4)\}$ is concordant if there is a permutation (i, j, k, l) of $(1, 2, 3, 4)$ such that

$$(x_i, x_j < x_k, x_l) \wedge [(y_i, y_j < y_k, y_l) \vee (y_i, y_j > y_k, y_l)]$$

and discordant if there is a permutation (i, j, k, l) of $(1, 2, 3, 4)$ such that

$$[(x_i, x_j < x_k, x_l) \vee (x_i, x_j > x_k, x_l)] \wedge [(y_i, y_k < y_j, y_l) \vee (y_i, y_k > y_j, y_l)]$$

It is straightforward to verify that

$$\begin{aligned} a(z_1, z_2, z_3, z_4) &= I(z_1, z_2 < z_3, z_4) + I(z_1, z_2 > z_3, z_4) \\ &\quad -I(z_1, z_3 < z_2, z_4) - I(z_1, z_3 > z_2, z_4) \end{aligned}$$

where I is the indicator function and $I(z_1, z_2 < z_3, z_4)$ is shorthand for $I(z_1 < z_3 \wedge z_1 < z_4 \wedge z_2 < z_3 \wedge z_2 < z_4)$. Hence,

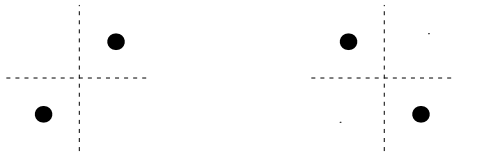
$$\begin{aligned} \tau^* &= 2P(X_1, X_2 < X_3, X_4 \wedge Y_1, Y_2 < Y_3, Y_4) + \\ &\quad 2P(X_1, X_2 < X_3, X_4 \wedge Y_1, Y_2 > Y_3, Y_4) - \\ &\quad 4P(X_1, X_2 < X_3, X_4 \wedge Y_1, Y_3 < Y_2, Y_4) \end{aligned} \tag{2}$$

Denoting the probability that four randomly chosen points are concordant as Π_{C_4} and the probability that they are discordant as Π_{D_4}, we obtain that the sum of the first two probabilities on the right hand side of (2) equal $\Pi_{C_4}/6$, while the last probability equals $\Pi_{D_4}/24$. Hence,

$$\tau^* = \frac{2\Pi_{C_4} - \Pi_{D_4}}{6} \tag{3}$$

The reason that Π_{C_4} is given twice as much weight as Π_{D_4} is related to there being twice as many discordant as concordant patterns in Figure 2. It can be seen that t^* and τ^* do not depend on the scale at which the variables are measured, but only on the ranks or grades of the observations. Four points are said to be *tied* if they are neither concordant nor discordant. Clearly, for continuous distributions the probability of tied observations is zero. Hence, under independence, when all configurations are equally likely, $\Pi_{C_4} = 1/3$ and $\Pi_{D_4} = 2/3$, and if one variable is a strictly monotone function of the other, then $\Pi_{C_4} = 1$ and $\Pi_{D_4} = 0$.

The definition of τ^* can easily be extended to X and Y in arbitrary metric spaces, but unfortunately Theorem 1 does not extend then, as it is possible that $\tau^* < 0$. This is shown by the



(a) Concordant pair (b) Discordant pair

Figure 1: Concordant and discordant pairs of points associated with Kendall's tau

following example. Consider a set of points $\{u_1, \ldots, u_8\} \subset \mathbf{R}^8$, where $u_i = (u_{i1}, \ldots, u_{i8})'$ such that $u_{ii} = 3$, $u_{ij} = -1$ if $i \neq j$ and $i, j \leq 4$ or $i, j \geq 5$, and $u_{ij} = 0$ otherwise. Suppose Y is uniformly distributed on $\{0, 1\}$, and given $Y = 0$, X is uniformly distributed on $u_1, \ldots, u_4$, and given $Y = 1$, X is uniformly distributed on $u_5, \ldots, u_8$. Then $\tau^* = -1/64$.

Still, for X and Y in arbitrary metric spaces, the following result suggests that a test that $\tau^* = 0$ against the alternative $\tau^* > 0$ can have power for certain interesting alternatives: if (X, Y) is the mixture of a non-degenerate independence model and a point mass, then $\tau^*(X, Y) > 0$ (Theorem 2 in Appendix A).

By the Cauchy-Schwarz inequality, the normalized value

$$\tau_b^* = \frac{\tau^*(X, Y)}{\sqrt{\tau^*(X, X)\tau^*(Y, Y)}}$$

does not exceed one. (Note that this notation is in line with Kendall's τ_b, defined analogously.)

Note that $\tau^*(X, Y)$ is a function of the copula, which is the joint distribution of $F_X(X)$ and $F_Y(Y)$, where F_X and F_Y are the cumulative distribution functions of X and Y. Nelsen (2006, Chapter 5) explores the way in which copulas can be used in the study of dependence between random variables, paying particular attention to Kendall's tau and Spearman's rho.

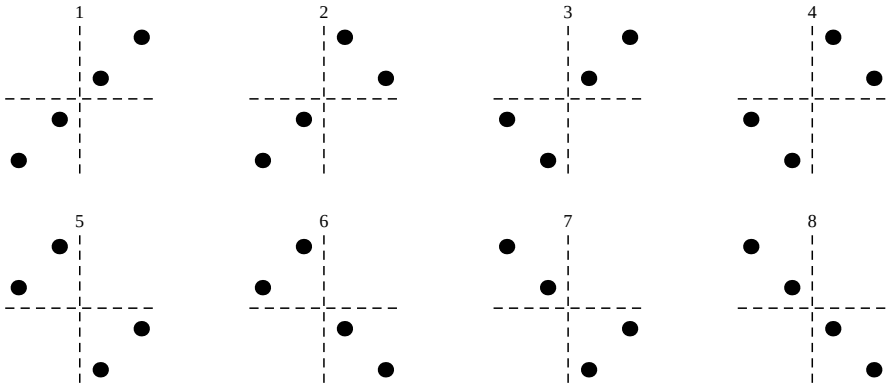
The remainder of the paper is organized as follows. In Section 3, a comparison is given with some other approaches in the literature. In particular, the Cramér von Mises test is essentially a special case and there are interesting similarities with a test devised by Hoeffding. In Section 4 we give a description of independence testing with an artificial and a real data example, briefly comparing the tests described in Section 3.2 with a test based on Kendall's tau and the chi-square test.

2 Proof of Theorem 1

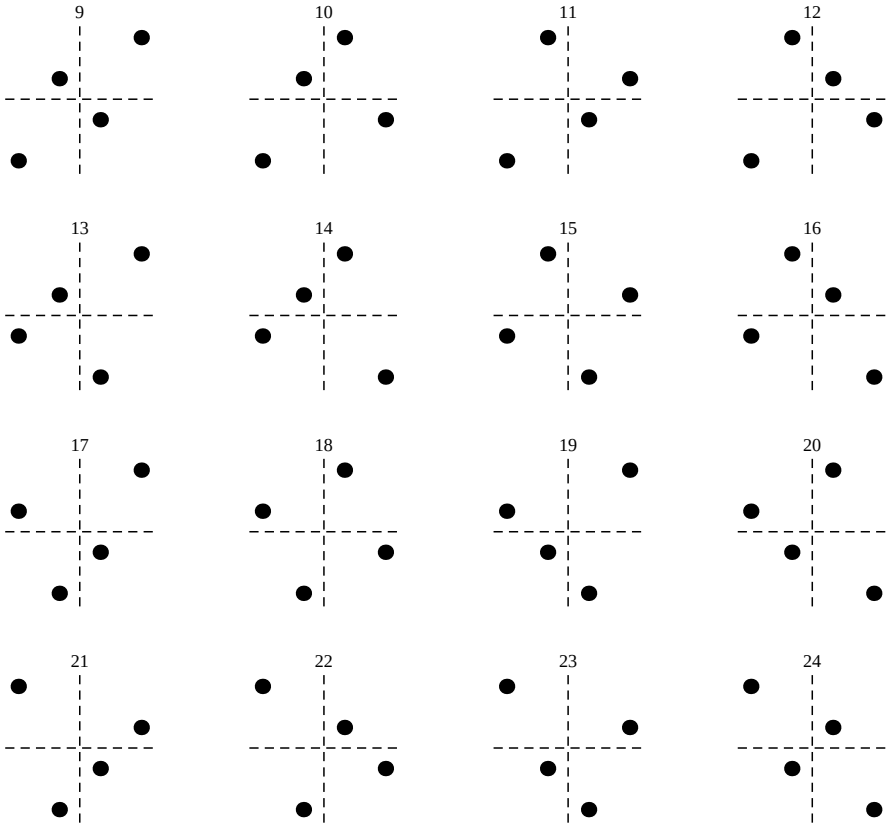
For the proof of Theorem 1, we need Lemma 1, which covers the case that one of the variables is binary. Note that Lemma 1 provides an extension of Theorem 1, which does not cover the discrete case.

Lemma 1 *If X is binary and Y given X is continuous or discrete, then $\tau^* \geq 0$ with equality if and only if X and Y are independent.*

Before giving the proof, let us first look at the form of τ^* when one of the variables is binary. Suppose $X \in \{0, 1\}$ and $Y \in \mathbf{R}$ and denote $U \equiv (Y|X = 0)$, $V \equiv (Y|X = 1)$, and $p = P(X = 0)$.



(a) Concordant quadruples



(b) Discordant quadruples

Figure 2: Configurations of concordant and discordant quadruples of points associated with τ^*. The dotted axes indicate strict separation of points in different quadrants; within a quadrant, no restrictions apply.

Then using (2) it is straightforward to verify that

$$\tau^* = 2p^2(1-p)^2\left[P(U_1,U_2<V_1,V_2)+P(V_1,V_2<U_1,U_2)-2P(U_1,V_1<U_2,V_2)\right] \tag{4}$$

Note that in this case, independence of X and Y is equivalent to U and V having identical distributions (see Section 3.1 for comments on the resulting two-sample test).

Below, we first prove Lemma 1 separately for the continuous and the discrete case, then we prove Theorem 1.

Proof of Lemma 1 (the continuous case): Continuity implies

$$P(U_1,U_2<V_1,V_2)+P(V_1,V_2<U_1,U_2)+4P(U_1,V_1<U_2,V_2)=1$$

so by (4), $\tau^* \geq 0$ reduces to

$$P(U_1,U_2<V_1,V_2)+P(V_1,V_2<U_1,U_2)\geq\frac{1}{3}.$$

Now, with G and H the distribution functions of U and V, respectively,

$$P\left(U_1,U_2<V_1,V_2\right)+P\left(V_1,V_2<U_1,U_2\right)=$$

$$2\int_{\mathbb{R}}G^2\left(1-H\right)dH+2\int_{\mathbb{R}}\left(1-G\right)^2HdH=$$

$$2\int_{\mathbb{R}}\left(G^2+H-2HG\right)dH=$$

$$2\int_{\mathbb{R}}\left(H-H^2\right)dH+2\int_{\mathbb{R}}\left(H-G\right)^2dH\geq\frac{1}{3}$$

with equality if and only if $H\equiv G$. □

Proof of Lemma 1 (the discrete case): The lemma is true for random variables that take two values each (by inspection).

Assume now it is true for random variables with k values. Let A_k be $P\left(U_1,U_2<V_1,V_2\right)+P\left(U_1,U_2>V_1,V_2\right)$ and B_k be $P\left(U_1,V_1>U_2,V_2\right)$ for some random variables with k values.

Suppose now that U_i^* has a mixed distribution: it has the same distribution as $U_i$with probability $1-\alpha$ and takes a different value larger than all previous possible values of U_i with probability α.

Similarly V_i^* has a mixed distribution: it takes the same distribution as V_i with probability $1-\beta$ and takes the different value mentioned before with probability β.

We now have

$$P\left(U_1^*,U_2^*<V_1^*,V_2^*\right)+P\left(U_1^*,U_2^*>V_1^*,V_2^*\right)=\alpha^2\beta^2A_k+\beta^2\left(1-\alpha^2\right)+\alpha^2\left(1-\beta^2\right)$$

and

$$P\left(U_1,V_1>U_2,V_2\right)=\alpha^2\beta^2B_k+\alpha\beta\left(1-\alpha\beta\right).$$

Hence

$$P\left(U_1^*,U_2^*<V_1^*,V_2^*\right)+P\left(U_1^*,U_2^*>V_1^*,V_2^*\right)-2P\left(U_1,V_1>U_2,V_2\right)=$$

$$\alpha^2\beta^2\left(A_k-2B_k\right)+\beta^2\left(1-\alpha^2\right)+\alpha^2\left(1-\beta^2\right)-2\alpha\beta\left(1-\alpha\beta\right)=$$

$$\alpha^2\beta^2\left(A_k - 2B_k\right) + \left(\alpha - \beta\right)^2.$$

But $A_k \geq 2B_k$ with equality if and only if the distributions are identical, so now for random variables taking $k+1$ values the statement is true with equality if and only if $A_k = 2B_k$ and $\alpha = \beta$, that is, when the distributions are identical. □

Proof of Theorem 1: We assume the distribution of (X_i, Y_i) is continuous. We can see that we need to prove that

$$P\left(Y_1, Y_2 < Y_3, Y_4 \& X_1, X_2 < X_3, X_4\right) + P\left(Y_1, Y_2 > Y_3, Y_4 \& X_1, X_2 < X_3, X_4\right) \geq$$

$$\frac{1}{3}P\left(X_1, X_2 < X_3, X_4\right).$$

This is because

$$P\left(Y_1, Y_2 < Y_3, Y_4 \& X_1, X_2 < X_3, X_4\right) + P\left(Y_1, Y_2 > Y_3, Y_4 \& X_1, X_2 < X_3, X_4\right) +$$

$$4P\left(Y_1, Y_3 > Y_2, Y_4 \& X_1, X_2 < X_3, X_4\right) = 1.$$

We now have that

$$P\left(Y_1, Y_2 < Y_3, Y_4 \& X_1, X_2 < X_3, X_4\right) + P\left(Y_1, Y_2 > Y_3, Y_4 \& X_1, X_2 < X_3, X_4\right) =$$

$$2\int\int\int_{\mathbb{R}^3} P\left(Y < y | X = x_1\right) P\left(Y < y | X = x_2\right)\left(1 - P\left(Y < y | X > x_1 \vee x_2\right)\right) \cdot$$

$$\left(P\left(X > x_1 \vee x_2\right)\right)^2 P\left(Y \in dy | X > x_1 \vee x_2\right) P\left(X \in dx_1\right) P\left(X \in dx_2\right) +$$

$$2\int\int\int_{\mathbb{R}^3} \left(1 - P\left(Y < y | X = x_1\right)\right)\left(1 - P\left(Y < y | X = x_2\right)\right) P\left(Y < y | X > x_1 \vee x_2\right) \cdot$$

$$\left(P\left(X > x_1 \vee x_2\right)\right)^2 P\left(Y \in dy | X > x_1 \vee x_2\right) P\left(X \in dx_1\right) P\left(X \in dx_2\right) =$$

$$2\int\int\int_{\mathbb{R}^3} \{P\left(Y < y | X = x_1\right) P\left(Y < y | X = x_2\right) + P\left(Y < y | X > x_1 \vee x_2\right)$$

$$-P\left(Y < y | X = x_1\right) P\left(Y < y | X > x_1 \vee x_2\right) - P\left(Y < y | X = x_2\right) P\left(Y < y | X > x_1 \vee x_2\right)\} \cdot$$

$$\left(P\left(X > x_1 \vee x_2\right)\right)^2 P\left(Y \in dy | X > x_1 \vee x_2\right) P\left(X \in dx_1\right) P\left(X \in dx_2\right) =$$

We now rewrite the quantity in brackets as

$$2\int\int\int_{\mathbb{R}^3} \left\{P\left(Y < y | X > x_1 \vee x_2\right) - \left(P\left(Y < y | X > x_1 \vee x_2\right)\right)^2\right\} \cdot$$

$$\left(P\left(X > x_1 \vee x_2\right)\right)^2 P\left(Y \in dy | X > x_1 \vee x_2\right) P\left(X \in dx_1\right) P\left(X \in dx_2\right) +$$

$$2\int\int\int_{\mathbb{R}^3} \left(P\left(Y < y | X > x_1 \vee x_2\right) - P\left(Y < y | X = x_1\right)\right)\left(P\left(Y < y | X > x_1 \vee x_2\right) - P\left(Y < y | X = x_2\right)\right) \cdot$$

$$\left(P\left(X > x_1 \vee x_2\right)\right)^2 P\left(Y \in dy | X > x_1 \vee x_2\right) P\left(X \in dx_1\right) P\left(X \in dx_2\right) \tag{5}$$

Now

$$\int_{\mathbb{R}} \left\{ P\left(Y < y | X > x_1 \vee x_2\right) - \left(P\left(Y < y | X > x_1 \vee x_2\right)\right)^2 \right\} P\left(Y \in dy | X > x_1 \vee x_2\right) = \frac{1}{6}$$

so the first of the two integrals in (5) is equal to

$$\frac{2}{6} \int \int_{\mathbb{R}^2} \left(P\left(X > x_1 \vee x_2\right)\right)^2 P\left(X \in dx_1\right) P\left(X \in dx_2\right) = \frac{1}{3} P\left(X_1, X_2 < X_3, X_4\right).$$

It remains to show that the second integral is non-negative (with 0 for independence). We have

$$\int \int \int_{\mathbb{R}^3} \left(P\left(Y < y | X > x_1 \vee x_2\right) - P\left(Y < y | X = x_1\right)\right) \left(P\left(Y < y | X > x_1 \vee x_2\right) - P\left(Y < y | X = x_2\right)\right) \cdot$$

$$\left(P\left(X > x_1 \vee x_2\right)\right)^2 P\left(X \in dx_1\right) P\left(X \in dx_2\right) P\left(Y \in dy | X > x_1 \vee x_2\right) =$$

$$\int \int \int_{\mathbb{R}^3} \left(P\left(Y < y, X > x_1 \vee x_2\right) P\left(X \in dx_1\right) - P\left(Y < y, X \in dx_1\right) P\left(X > x_1 \vee x_2\right)\right) \cdot$$

$$\left(P\left(Y < y, X > x_1 \vee x_2\right) P\left(X \in dx_2\right) - P\left(Y < y, X \in dx_2\right) P\left(X > x_1 \vee x_2\right)\right) P\left(Y \in dy | X > x_1 \vee x_2\right).$$

The integrand of the expression above has the same sign as

$$\int \int_{\mathbb{R}^2} \left(P\left(X > x_1 \vee x_2 | Y < y\right) P\left(X \in dx_1\right) - P\left(X \in dx_1 | Y < y\right) P\left(X > x_1 \vee x_2\right)\right) \cdot$$

$$\left(P\left(X > x_1 \vee x_2 | Y < y\right) P\left(X \in dx_2\right) - P\left(X \in dx_2 | Y < y\right) P\left(X > x_1 \vee x_2\right)\right) \frac{P\left(X > x_1 \vee x_2 | Y = y\right)}{P\left(X > x_1 \vee x_2\right)}. \quad (6)$$

To simplify notation, we now define $P\left(X < x\right) = G\left(x\right)$, $P\left(X > x\right) = \overline{G}\left(x\right)$, $P\left(X \in dx\right) = g\left(x\right) dx$, $P\left(X < x | Y < y\right) = H\left(x\right)$, $P\left(X > x | Y \leq y\right) = \overline{H}\left(x\right)$, $P\left(X \in dx | Y < y\right) = h\left(x\right) dx$, $P\left(X < x | Y = y\right) = F\left(x\right)$, $P\left(X > x | Y = y\right) = \overline{F}\left(x\right)$ and $P\left(X \in dx | Y = y\right) = f\left(x\right) dx$. We rewrite (6) as

$$\int \int_{\mathbb{R}^2} \left(\overline{H}\left(x_1 \vee x_2\right) g\left(x_1\right) - \overline{G}\left(x_1 \vee x_2\right) h\left(x_1\right)\right) \cdot$$

$$\left(\overline{H}\left(x_1 \vee x_2\right) g\left(x_2\right) - \overline{G}\left(x_1 \vee x_2\right) h\left(x_2\right)\right) \frac{\overline{F}\left(x_1 \vee x_2\right)}{\overline{G}\left(x_1 \vee x_2\right)} dx_1 dx_2 =$$

$$2 \int_{\mathbb{R}} \int_{-\infty}^{x_2} \left(\overline{H}\left(x_2\right) g\left(x_1\right) - \overline{G}\left(x_2\right) h\left(x_1\right)\right) dx_1 \left(\overline{H}\left(x_2\right) g\left(x_2\right) - \overline{G}\left(x_2\right) h\left(x_2\right)\right) \frac{\overline{F}\left(x_2\right)}{\overline{G}\left(x_2\right)} dx_2 =$$

$$2 \int_{\mathbb{R}} \left(\overline{H}\left(x_2\right) G\left(x_2\right) - \overline{G}\left(x_2\right) H\left(x_2\right)\right) \left(\overline{H}\left(x_2\right) g\left(x_2\right) - \overline{G}\left(x_2\right) h\left(x_2\right)\right) \frac{\overline{F}\left(x_2\right)}{\overline{G}\left(x_2\right)} dx_2 =$$

$$2 \int_{\mathbb{R}} \left(G\left(x_2\right) - H\left(x_2\right)\right) \left(\overline{H}\left(x_2\right) g\left(x_2\right) - \overline{G}\left(x_2\right) h\left(x_2\right)\right) \frac{\overline{F}\left(x_2\right)}{\overline{G}\left(x_2\right)} dx_2 =$$

$$2 \int_{\mathbb{R}} \left(G\left(x_2\right) - H\left(x_2\right)\right) \left(\overline{H}\left(x_2\right) g\left(x_2\right) - \overline{G}\left(x_2\right) g\left(x_2\right) + \overline{G}\left(x_2\right) g\left(x_2\right) - \overline{G}\left(x_2\right) h\left(x_2\right)\right) \frac{\overline{F}\left(x_2\right)}{\overline{G}\left(x_2\right)} dx_2 =$$

$$2\int_{\mathbb{R}}\left(G\left(x_2\right)-H\left(x_2\right)\right)\left(\overline{H}\left(x_2\right)-\overline{G}\left(x_2\right)\right)g\left(x_2\right)\frac{\overline{F}\left(x_2\right)}{\overline{G}\left(x_2\right)}dx_2+$$
$$2\int_{\mathbb{R}}\left(G\left(x_2\right)-H\left(x_2\right)\right)\left(g\left(x_2\right)-h\left(x_2\right)\right)\overline{G}\left(x_2\right)\frac{\overline{F}\left(x_2\right)}{\overline{G}\left(x_2\right)}dx_2=$$
$$2\int_{\mathbb{R}}\left(G\left(x_2\right)-H\left(x_2\right)\right)^2 g\left(x_2\right)\frac{\overline{F}\left(x_2\right)}{\overline{G}\left(x_2\right)}dx_2+$$
$$2\int_{\mathbb{R}}\left(G\left(x_2\right)-H\left(x_2\right)\right)\left(g\left(x_2\right)-h\left(x_2\right)\right)\int_{x_2}^{\infty}f\left(z\right)dzdx_2=$$
$$2\int_{\mathbb{R}}\left(G\left(x_2\right)-H\left(x_2\right)\right)^2 g\left(x_2\right)\frac{\overline{F}\left(x_2\right)}{\overline{G}\left(x_2\right)}dx_2+$$
$$2\int_{\mathbb{R}}\int_{-\infty}^{z}\left(G\left(x_2\right)-H\left(x_2\right)\right)\left(g\left(x_2\right)-h\left(x_2\right)\right)dx_2 f\left(z\right)dz=$$
$$2\int_{\mathbb{R}}\left(G\left(x_2\right)-H\left(x_2\right)\right)^2 g\left(x_2\right)\frac{\overline{F}\left(x_2\right)}{\overline{G}\left(x_2\right)}dx_2+\int_{\mathbb{R}}\left(G\left(z\right)-H\left(z\right)\right)^2 f\left(z\right)dz$$

which is non-negative and can only be 0 if $G(x) = H(x)$ for all x, that is if $P(X < x) = P(X < x|Y < y)$ for (almost) all x and y, which is equivalent to independence. □

3 Comparison to other approaches

If one of the variables is binary, our approach leads to the Cramér von Mises test, as described in Section 3.1. In Sections 3.2 and 3.3 the two direct competitors to tests based on τ^* known to the authors, one originally by Hoeffding (1948) and another originally by De Wet (1980) and Deheuvels (1981), are discussed. Both these competitors share the property of our approach that they are rank-based and consistent for all alternatives.

Useful and extensive discussions of other ordinal data and nonparametric methods for independence testing are given Agresti (2010), Hollander and Wolfe (1999) and Sheskin (2007).

3.1 The two-sample case and relation to the Cramér von Mises test

The two-sample Cramér von Mises test is used to test whether or not two samples are drawn from the same distribution, and is consistent for any alternative. We show that if one of the variables is binary and the conditional distribution of the other variable is continuous, a test based on τ^* coincides with the Cramér von Mises test. We argue that the test based on τ^* has a possible advantage for discrete distributions.

We now give the relationship with the Cramér von Mises test. Let G be the distribution function of U and let H be the distribution function of V. With $F_\alpha = \alpha G + (1-\alpha)H$ let

$$C_\alpha = \int (G-H)^2 dF_\alpha \tag{7}$$

Then C_α is zero if and only if $G = H$, i.e., if and only if X and Y are independent. The Cramér von Mises test statistic is based on an estimate of C_p. First, note:

Lemma 2 *For $\alpha \in \mathbf{R}$, C_α does not depend on α.*

Proof: The lemma is implied because

$$\int (G-H)^2 dH - \int (G-H)^2 dG = \int (G-H)^2 d(G-H) = \left[\frac{1}{3}(G-H)^3\right]_{-\infty}^{\infty} = 0$$

□

The relationship between the Cramér von Mises test, which is based on C_p, and τ^* is given by the following:

Lemma 3 *If X is binary and Y given X is continuous, then $\tau^* = 6p^2(1-p)^2 C_p$.*

Proof: First note that

$$\int H dH = \frac{1}{2} \quad \text{and} \quad \int H^2 dH = \frac{1}{3} \tag{8}$$

Now continuity implies

$$P(U_1, U_2 < V_1, V_2) + P(V_1, V_2 < U_1, U_2) + 4P(U_1, V_1 < U_2, V_2) = 1$$

Hence by (4) and (8),

$$\begin{aligned}
\tau^* &= 2p^2(1-p)^2 \left[\frac{3}{2}P(U_1, U_2 < V_1, V_2) + \frac{3}{2}P(V_1, V_2 < U_1, U_2) - \frac{1}{2}\right] \\
&= 2p^2(1-p)^2 \left[3\int G^2 H dH + 3\int (1-G)^2 H dH - \frac{1}{2}\right] \\
&= 2p^2(1-p)^2 \left[3\int (G^2 - 2GH + H) dH - \frac{1}{2}\right] \\
&= 2p^2(1-p)^2 \left[3\int (G^2 - 2GH + H^2) dH\right] \\
&= 2p^2(1-p)^2 \left[3\int (G-H)^2 dH\right]
\end{aligned}$$

The lemma now follows from Lemma 2 □

Note that, for discrete distributions, the definition of C_p unsatisfactorily depends on the way G and H are defined, e.g., whether we define $G(u) = P(U < u)$ or $G(u) = P(U \leq u)$. Since τ^* deals naturally with discreteness of random variables, tests based on τ^* might serve as an alternative for the Cramér von Mises test if discreteness is present.

3.2 Hoeffding's H

Hoeffding's (1948) coefficient for measuring deviation from independence for a bivariate distribution function F_{12} with marginal distribution functions F_1 and F_2 is defined as

$$H = \int (F_{12} - F_1 F_2)^2 dF_{12}$$

(See also Blum, Kiefer, & Rosenblatt, 1961; Hollander & Wolfe, 1999 and Wilding & Mudholkar, 2008.) An alternative formulation given by Hoeffding is

$$H = \frac{1}{4} E\phi((X_1, X_2, X_3)\phi((X_1, X_4, X_5)\phi((Y_1, Y_2, Y_3)\phi((Y_1, Y_4, Y_5)$$

where $\phi(z_1, z_2, z_3) = I(z_1 \geq z_2) - I(z_1 \geq z_3)$.

Interestingly, Hoeffding's H has an interpretation in terms of concordance and discordance probabilities closely related to the interpretation of τ^*. With

$$\begin{aligned}
F_{12}(x,y) &= P(X \leq x, Y \leq y) \\
F_{1\overline{2}}(x,y) &= P(X \leq x, Y > y) = F_1(x) - F_{12}(x,y) \\
F_{\overline{1}2}(x,y) &= P(X > x, Y \leq y) = F_2(y) - F_{12}(x,y) \\
F_{\overline{12}}(x,y) &= P(X > x, Y > y) = 1 - F_1(x) - F_2(y) + F_{12}(x,y)
\end{aligned}$$

we have the equality

$$F_{12} - F_1 F_2 = F_{12} F_{\overline{12}} - F_{1\overline{2}} F_{\overline{1}2} \tag{9}$$

Let five points be H-concordant if four are configured as in Figure 2(a) and the fifth is on the point where the axes cross and, analogously, five points are H-discordant if four are configured as in Figure 2(b) and the fifth is on the point where the axes cross. Denote the probabilities of H-concordance and discordance by Π_{C_5} and Π_{D_5}. Then

$$\int \left(F_{12}^2 F_{\overline{12}}^2 + F_{1\overline{2}}^2 F_{\overline{1}2}^2\right) dF_{12} = \frac{2!2!}{5!}\Pi_{C_5} = \frac{1}{30}\Pi_{C_5}$$

and

$$\int F_{12} F_{1\overline{2}} F_{\overline{1}2} F_{\overline{12}} dF_{12} = \frac{1}{5!}\Pi_{D_5} = \frac{1}{120}\Pi_{D_5}$$

Hence, using (9),

$$H = \int \left(F_{12} F_{\overline{12}} - F_{1\overline{2}} F_{\overline{1}2}\right)^2 dF_{12} = \frac{2\Pi_{C_5} - \Pi_{D_5}}{60}$$

3.3 De Wet and Deheuvels' D

A coefficient related to Hoeffding's H is

$$D = \int (F_{12} - F_1 F_2)^2 dF_1 F_2$$

Tests based on estimators of this coefficient were studied by De Wet (1980) and Deheuvels (1981). Bergsma (2006) showed that in the continuous case, with

$$h(z_1, z_2, z_3, z_4) = |z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|,$$

$$D = Eh(F_1(X_1), F_1(X_2), F_1(X_3), F_1(X_4))h(F_2(Y_1), F_2(Y_2), F_2(Y_3), F_2(Y_4))$$

The latter definition is suitable for discontinuous X and Y as well and has the advantage that it does not depend on the way F_{12} is defined.

		Y						
		1	2	3	4	5	6	7
X	1	2	1	0	0	0	1	2
	2	1	2	0	0	0	2	1
	3	0	0	2	1	2	0	0
	4	0	0	1	1	1	0	0
	5	0	0	1	2	1	0	0

Table 1: Artificial contingency table containing multinomial counts. Kendall's tau and the chi-square test do not yield a significant association, but a permutation test based t^2 yields $p = 0.035$

3.4 Comparison of τ^*, H, and D

Hoeffding's H is more complex than τ^* in that it is based on concordance and discordance of five points rather than four. See also Kruskal (1958), who compares Kendall's tau and Spearman's rho which are based on concordance and discordance probabilities of two and three points respectively and for this reason expresses some tentative preference for the simpler Kendall's tau. A possible drawback of D is that there seems to be some arbitrariness in the use of rank scores; for example, one might also use normal scores. Of the three coefficients, τ^* may thus have some advantage. Tests based on τ^* are discussed in the next section, which includes a remark on power.

4 Testing independence

A suitable test for independence is a permutation test which rejects the independence hypothesis for large values of t^*. For every permutation π of the observed y-values, the sample τ^*-value t^*_π is computed, and the p-value is the proportion of the $\{t^*_\pi\}$ which exceed t^*. As is well-known, the permutation test conditions on the empirical marginal distributions, which are sufficient statistics for the independence model. In categorical data analysis, it is usually referred to as an exact conditional test. In practice, the number of permutations may be too large to compute and a random sample of permutations is taken, which is also called a resampling test. Note that there doesn't seem to be a need for an asymptotic approximation to the sampling distribution of t^*.

Direct evaluation of t^* requires computational time $O(n^4)$, which may be practically infeasible for moderately large samples, but t^* can be well-approximated by taken a random sample of subsets of four observations. The proof of Theorem 1 suggests that the complexity can be reduced to $O(n^3)$. An open problem is what the minimum computational complexity of computing t^* is.

Below, we compare various tests of independence using an artificial and a real data set.

An artificial multinomial table of counts is given in Table 1, where X and Y are ordinal variables with 5 and 7 categories. Visually, we can detect an association pattern, but as it is non-monotonic a test based on Kendall's tau does not yield a significant p-value. The chi-square test also yields a non-significant $p = 0.253$ (based on 10^6 resamples), while a permutation test based on t^* yields $p = 0.035$ (10^4 resamples), giving evidence of an association. We also did tests based on D, which yields $p = 0.045$ (10^4 resamples), and the test based on Hoeffding's H yields $p = 0.028$ (4000 resamples). In this example, using a test designed for ordinal data with power against broad alternatives, evidence for an association can be found, which is not possible with a nominal data test like the chi-square test or with a test based on Kendall's tau.

		Change in size of Ulcer Crater (Y)			
		Larger	Healed ($< \frac{2}{3}$)	Healed ($\geq \frac{2}{3}$)	Healed
Treatment group (X)	A	6	4	10	12
	B	11	8	8	5

Table 2: Results of study comparing two treatments of gastric ulcer

Table 2 shows data from a randomized study to compare two treatments for a gastric ulcer crater, and was previously analyzed in Agresti (2010). Using 10^5 resamples, the chi-square test yields $p = 0.118$, Kendall's tau yields $p = 0.019$, t^* yields $p = 0.028$, D yields $p = 0.026$, and using 10^4 resamples Hoeffding's H yields $p = 0.006$.

For future research, more understanding is needed concerning the power of the tests based on t^*, H, D, and the chi-square test. We have done some limited simulations and, unsurprisingly, found that for all four there are alternatives for which they are most powerful. However, so far we were unable to detect any patterns which can lead to useful advice on when to use which test; it appears we need a better understanding than we currently appear to have of the types of alternatives that might be of most interest.

A Mixing an independence model with a point mass

Let Ω_1 and Ω_2 be metric spaces and suppose $X \in \Omega_1$ and $Y \in \Omega_2$ are independent non-degenerate random variables. Consider the mixture of (X, Y) with the degenerate random variable on the point $(x_0, y_0) \in \Omega_1 \times \Omega_2$, that is, for some $0 < p < 1$ the mixture (X', Y') is defined as

$$(X', Y') = \begin{cases} (X, Y) & \text{with probability } p \\ (x_0, y_0) & \text{with probability } 1 - p \end{cases}$$

Then

Theorem 2 $\tau^*(X', Y') > 0$.

Proof: The proof is done by conditioning on the number of occurrences of (x_0, y_0) among the iid $(X'_1, Y'_1), \ldots, (X'_4, Y'_4)$. Clearly, (x_0, y_0) can occur 0 to 4 times, each with positive probability, and τ^* is the sum of these probabilities times the conditional expectations of the product of

$$\text{sign}(|X'_1 - X'_2| + |X'_3 - X'_4| - |X'_1 - X'_3| - |X'_2 - X'_4|) \tag{10}$$

and

$$\text{sign}(|Y'_1 - Y'_2| + |Y'_3 - Y'_4| - |Y'_1 - Y'_3| - |Y'_2 - Y'_4|) \tag{11}$$

Conditionally on the number of occurrences of (x_0, y_0), the expectation of the product of (10) and (11) equals the product of their expectations. If (x_0, y_0) occurs 3 or 4 times, both (10) and (11) are zero, hence zero is contributed to τ^*. Conditionally on (x_0, y_0) occurring 0 or 1 times, the expectations of both (10) and (11) can easily be seen to equal zero by symmetry reasons.

To prove the theorem, it remains to be shown that conditionally on (x_0, y_0) occurring twice, both (10) and (11) have positive expectation. If either (X_1, Y_1) and (X_4, Y_4) or (X_2, Y_2) and

(X_3, Y_3) equal (x_0, y_0), both (10) and (11) are zero and this will not contribute to τ^*. Without loss of generality we now only need to consider (X_3, Y_3) and (X_4, Y_4) equalling (x_0, y_0). Then (10) reduces to

$$\text{sign}(|X_1 - x_0| + |X_2 - x_0| - |X_1 - X_2|)$$

and (11) reduces to

$$\text{sign}(|Y_1 - y_0| + |Y_2 - y_0| - |Y_1 - Y_2|)$$

By the triangle inequality, both are nonnegative. Since the X_i and Y_i are non-degenerate, both have positive probability of being positive and so have positive expectations. Hence $\tau^* > 0$. □

B Mathematica programme for verifying $\tau^* \geq 0$ for 3×3 table

The code below verifies nonnegativity of τ^* for 3×3 contingency tables with uniform marginals. In the code, we replaced the uniform marginals by a variety of other marginals and always obtained the same result.

```
(*module to compute tau-star, where p is an r x c table of probabilities,
x and y are r x 1 and c x 1 vectors of scores*)
taustarCD[p_, x_, y_] := Module[{sgn, r, c, sxy},
  sgn = Compile[{{a,_Integer},{b,_Integer},{c,_Integer},{d,_Integer}},
     Sign[Abs[a - b] + Abs[c - d] - Abs[a - c] - Abs[b - d]]];
  {r, c} = {Length[x], Length[y]};
  sxy = Sum[p[[i1,j1]]p[[i2,j2]]p[[i3,j3]]p[[i4, j4]]
      sgn[x[[i1]],x[[i2]],x[[i3]],x[[i4]]]
      sgn[y[[j1]],y[[j2]],y[[j3]],y[[j4]]],
      {i1,r},{i2,r},{i3,r},{i4,r},{j1,c},{j2,c},{j3,c},{j4,c}]
  ]

{r, c} = {3, 3};(*3 rows, 3 columns*)

(*fixed uniform marginals, can be modified*)
xmarg = {1/3, 1/3, 1/3};
ymarg = {1/3, 1/3, 1/3};

(*compute t: tau-star for 3x3 table*)
pp = Table[p[i, j], {i, r}, {j, c}];
x = Range[r];
y = Range[c];
t = taustarCD[pp, x, y] // Simplify // PowerExpand // Simplify;

(*specify assumption of nonnegative probabilities*)
nonnegprob = Map[# > 0 &, Flatten[pp]];
```

```
(*define t2 which is t but a function of only (r-1)(c-1)
probabilities with given marginals*)
fixmarg1 = Table[p[i,c]->xmarg[[i]]-Sum[p[i,j],{j,1,c-1}],{i,1,r}];
fixmarg2 = Table[p[r,j]->ymarg[[j]]-Sum[p[i,j],{i,1,r-1}],{j,1,c}];
t2 = t //. Join[fixmarg1, fixmarg2] // Simplify

(*check if t2>=0*)
Simplify[t2 >= 0, Assumptions -> nonnegprob]
```

Evaluation gives the result `True`.

References

Agresti, A. (2010). *Analysis of ordinal categorical data (second edition).* New York: Wiley.

Bergsma, W. P. (2006). A new correlation coefficient, its orthogonal decomposition, and associated tests of independence. *arXiv:math/0604627v1 [math.ST].*

Blum, J. R., Kiefer, J., & Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The annals of mathematical statistics*, *32*, 485-498.

De Wet, T. (1980). Cramér-von Mises tests for independence. *J. Multivariate Anal.*, *10*, 38-50.

Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *J. Multivariate Anal.*, *11*, 102-113.

Hoeffding, W. (1948). A non-parametric test of independence. *Annals of Mathematical Statistics*, *19*, 546-557.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods.* NY: Wiley.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods.* New York: Oxford University Press.

Kruskal, W. H. (1958). Ordinal measures of association. *J.Am.Stat.Ass*, *53*, 814-861.

Nelsen, R. B. (2006). *An introduction to copulas.* New York: Springer.

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures (fourth edition).* Boca Raton: Chapman and Hall.

Wilding, G. E., & Mudholkar, G. S. (2008). Empirical approximations for Hoeffding's test of bivariate independence using two Weibull extensions. *Stat. Methodol.*, *5*(2), 160–170.